Chapter 18:

Binomial Regression Modeling

Ned Levine
Ned Levine & Associates
Houston, TX

Dominique Lord
Zachry Dept. of
Civil Engineering
Texas A & M University
College Station, TX

Byung-Jung Park
Korea Transport Institute
Goyang, South Korea

Table of Contents

Chapter 18:
Binomial Regression Modeling

Introduction

In this chapter, we discuss binomial regression models as applied to ungrouped data. Users should be familiar with the materials in Chapter 15, 16, and 17 before attempting to read this chapter. A good background in statistics is necessary to understand the material.

These are models that are applied to individual cases (records) and where the dependent variable has only two responses, expressed as 0 and 1. They are part of a family of regression models called *limited dependent variables* where the range of possible values is restricted. They are sometimes called *restricted dependent variables* or, if the restriction is one side of the distribution only, *censored dependent variables* or even *truncated dependent variables*. In chapters16 and 17, we discussed the Poisson family of regression models. This is a limited dependent variable in that 0 is the minimum since the Poisson models counts (i.e., for which the minimum number is 0).

However, with binomial regression models, the limitations are on both sides of the distribution, namely a minimum value of 0 and a maximum value of 1. Such a model is useful when there is a discrete choice between two alternatives, for example 'yes' versus 'no' on a survey or 'males' versus 'females' as a demographic distinction or even 'under age 65' versus '65 or older' for an age group distinction. The key is that there can only be two alternatives and that they have to be identified as '1' or '0'.

The underlying model is that of a probability, which also varies from 0 to 1. The problem, however, is that with a binomial variable, the underlying probabilities are not measured but only inferred from a discrete, binomial choice. Thus, the models that have been proposed estimate the underlying probability using only the two alternative values for the dependent variable.

The two models that we will examine are the logistic (usually called logit) model and the probit model, the two most common forms for estimating the underlying probabilities. Binomial functions are also the basic building block for discrete choice models that comprise models for estimating probabilities when there are more than two alternatives. These will be discussed in chapters 21 and 22.

Generalized Linear Models

The Generalized Linear Model (GLM) is a family of functions for estimating the relationship of many functions to a set of linear predictors in a regression framework (Liao, 1994; McCullagh & Nelder, 1989). It relates the expected mean of a distribution, μ, to a *link function*, η, which, in turn, is related to a set of linear predictors,

$$\eta_i = \beta_0 + \sum_1^K \beta_K X_{iK} + \epsilon_i \quad (18.1)$$

where, for case i, β_0 is the intercept, β_K is the coefficient of each of the *K* independent variables, X_{iK}, and ε_i is an error term. The coefficients are applied to individual records, i. To simplify notation, we will drop the case letter but it will be understood that the parameters apply to individual cases.

Not all functions can be estimated this way, essentially only those that belong to the exponential family of functions and which have a concave, closed-form solution. In the classic linear form of the GLM model (Ordinary Least Squares, or OLS), which we examined in chapter 15, the link function is simply the mean itself,

$$\eta = \mu \quad (18.2)$$

In the Poisson form, which we examined in Chapters 16 and 17, the link function is the natural log of the mean,

$$\eta = Ln(\mu) \quad (18.3)$$

This brings us to binomial regression and the two forms which are also part of the GLM family. First, there is the ***logistic*** **(**or ***logit*****)** model where the link function is related to the *log of the odds* ,

$$\eta = Ln[(\mu/(1-\mu)] \quad (18.4)$$

Second, there is the ***probit*** model where the link function is related to the inverse of the standard normal cumulative distribution,

$$\eta = \Phi^{-1}(\mu) \quad (18.5)$$

There are other link functions that can be expressed by the GLM model, but we will concentrate on the logit and probit models. The logit is the most common way to relate a binomial outcome to a set of independent predictors with the probit used less often. In practice,

the logit and probit models produce more or less the same results (Greene, 2008). They differ primarily in the tails of the distribution with the probit approaching the limiting ends of the probability more quickly than the logit (Chen & Tsurumi, 2011; Hahn & Soyer, 2005).

Logistic Model

Logit

The logistic model is related to the binomial probability. It is usually called a logit model because it takes the log of the odds (*logit* and *log of the odds* are equivalent terms). If an event has two possible outcomes expressed as 0 and 1 (e.g. 'head' or 'tails', 'males' or 'females', 'A' or 'B', or any other binomial alternative), then its probability can be estimated for successive independent outcomes from *N* observations. Let *p* be the probability of obtaining one of the outcomes which takes the value 1 (call it A) with *1-p* (sometimes called *q*) being the probability of obtaining the outcome that takes the value 0 (call it B).

Binomial Distribution

The binomial distribution defines the distribution of alternative A in *O* successive samples by (Wikipedia, 2011a; Hosmer & Lemeshow, 2001):

$$P(Y = O) = \binom{N}{O} p^K (1-p)^{(N-K)} \tag{18.6}$$

where *P(Y=O)* is the probability of obtaining exactly *O* instances from *N* observations, *p* is the probability of obtaining A for one observation, and $\binom{N}{O}$ is the number of *combinations* for getting exactly *O* outcomes for A and *N-O* outcomes for B, and is expressed by

$$\binom{N}{K} = \frac{N!}{O!(N-O)!} = \frac{N(N-1)(N-2)\ldots(1)}{[O(O-1)(O-2)\ldots1][(N-O)(N-O-1)(N-O-2)\ldots1]} \tag{18.7}$$

where *!* is a factorial.

The probability is always estimated with respect to A (or the probability of achieving a 1). For example, if *p* for A is 0.4 (and, therefore, the probability for B is *1-p*, or 0.6) and there are 10 successive observations, each of which is independent, the probability of getting exactly 4 instances of *p* and 6 instances of *(1-p)* is:

$$\binom{100}{40} = \frac{10!}{4!(6)!} = \frac{(10)(9)(8)\ldots(1)}{[(4)(3)(2)(1)][(6)(5)(4)(3)(2)*(1)]} = (210)(.4)^4(.6)^6 = 0.2508$$

The probability is often called a Bernoulli trial, named after the Swiss mathematician Jacob Bernoulli (1654-1705; Wikipedia, 2011b; Hosmer & Lemeshow, 2001). Notice that the successive outcomes (sometimes called 'trials' or 'experiments') must be independent. That is, the probability of achieving either of the two outcomes in an observation (or trial) must be constant across observations and unrelated to prior observations. That is, the outcomes are random and independent. The assumption of independence of each observation (or trial or experiment) is different from the MCMC method that we discussed in Chapter 17 where the results of each sample depend on the value from the previous sample.

In a binomial experiment, there are exactly *N* observations and the function P(X=K) is called the *binomial distribution*. The binomial distribution, in turn, is a special case of the *Poisson distribution* which is a sum of *N* independent Bernoulli trials with a constant probability for each choice. The Poisson distribution expresses the probability of a given number of events occurring (in time or in space) if these events are independent and occur with a known probability. In other words, the Poisson distribution, which we examined in Chapters 16 and 17, is a more general case of the binomial distribution and, in turn, is part of the GLM family of models. The binomial distribution becomes the Poisson for very large samples (i.e., as N approaches infinity) and when p is very small (Lord, Washington, & Ivan, 2005).

Odds Ratio

Another way to look at the probability of obtaining alternative A compared to alternative B is through the *Odds Ratio* (or just Odds). This is the ratio of *p* to *1-p*, or

$$Odds\ ratio = \frac{p}{1-p} \tag{18.8}$$

and expresses the relative likelihood of obtaining outcome A relative to outcome B. For example, if p is 0.4 then 1-p is 0.6 and the odds ratio is 0.4/0.6 = 0.667. Alternatively, if p is 0.7 and 1-p is 0.3, then the odds ratio is 0.7/0.3 = 2.33. Finally, if p and 1-p are equal (i.e., both are 0.5), then the odds ratio is 0.5/0.5 = 1. Note that with the odds ratio, a value greater than 1 indicates that A is more likely to occur than B while a value less than 1 indicates that A is less likely to occur than B (or, conversely, B is more likely to occur than A). Thus, this means that A is about 2.3 times more likely to occur than B in the example.

Log of the Odds Ratio

Since the logit is the natural log of the odds ratio, if we let the probability, *p*, represent an estimate of the mean of the function, μ, then the logit model relates the logit of *p* to a linear set of predictors,

$$\eta = \text{Ln}\left(\frac{p}{1-p}\right) = \beta_0 + \sum_1^K \beta_K X_K + \varepsilon \quad (18.9)$$

This link function does three things that are useful. First, it relates the probability of a binomial outcome to a set of linear predictors. Second, taking the exponent of the logit relates the odds ratio to a set of linear predictors,

$$\left(\frac{p}{1-p}\right) = e^{\beta_0 + \sum_1^K \beta_K X_K + \varepsilon} \quad (18.10)$$

Therefore, the relative probability of obtaining outcome A relative to outcome B can be expressed as an exponential function of a linear set of predictors. This means that one can relate the odds ratio to a set of predictors that can account for the likelihood of A relative to that of B. Comparisons can then be made and linked to other variable. For example, suppose we categorize weapon use by robbers into two categories: 1) gun, knife or other weapon, and 2) using bodily force or threat. Then, the probability of using a physical weapon relative to bodily force can be expressed as a function of one or more independent variables.

Third, by taking the log of the odds ratio, the dependent variable is now a continuous variable that varies from minus infinity to plus infinity (though in practice between -3 and +3). In other words, the logit also eliminates the range restriction of a dependent binomial variable since the logit can have any value between minus and plus infinity.

Figure 18.1 shows the effect of transforming a probability into a logit. Notice how the function is fairly flat from about 0.2 to 0.8 beyond which the logit accelerates. When we reverse the axes and plot the effect of a logit on the probability, we have the classic S-shaped curve (Figure 18.2). The effect of a change in the logit on the probability is most pronounced in the middle of the probability range whereas there is less change at the low and high ends of the logit. In other words, the effect of the logit is to linearize the probability within the middle range of probability in order to allow a regression model to be tested.

Logistic Form

Equation 18.9 expresses the log of the odds as a function of linear predictors. Manipulating equation 18.9 leads to a solution for *p*,

$$P(Y=1) = \frac{e^{\beta_0 + \sum_1^K \beta_K X_K}}{1 + e^{\beta_0 + \sum_1^K \beta_K X_K}} = \frac{1}{1 + e^{-(\beta_0 + \sum_1^K \beta_I X_K)}} \quad (18.11)$$

Figure 18.1:

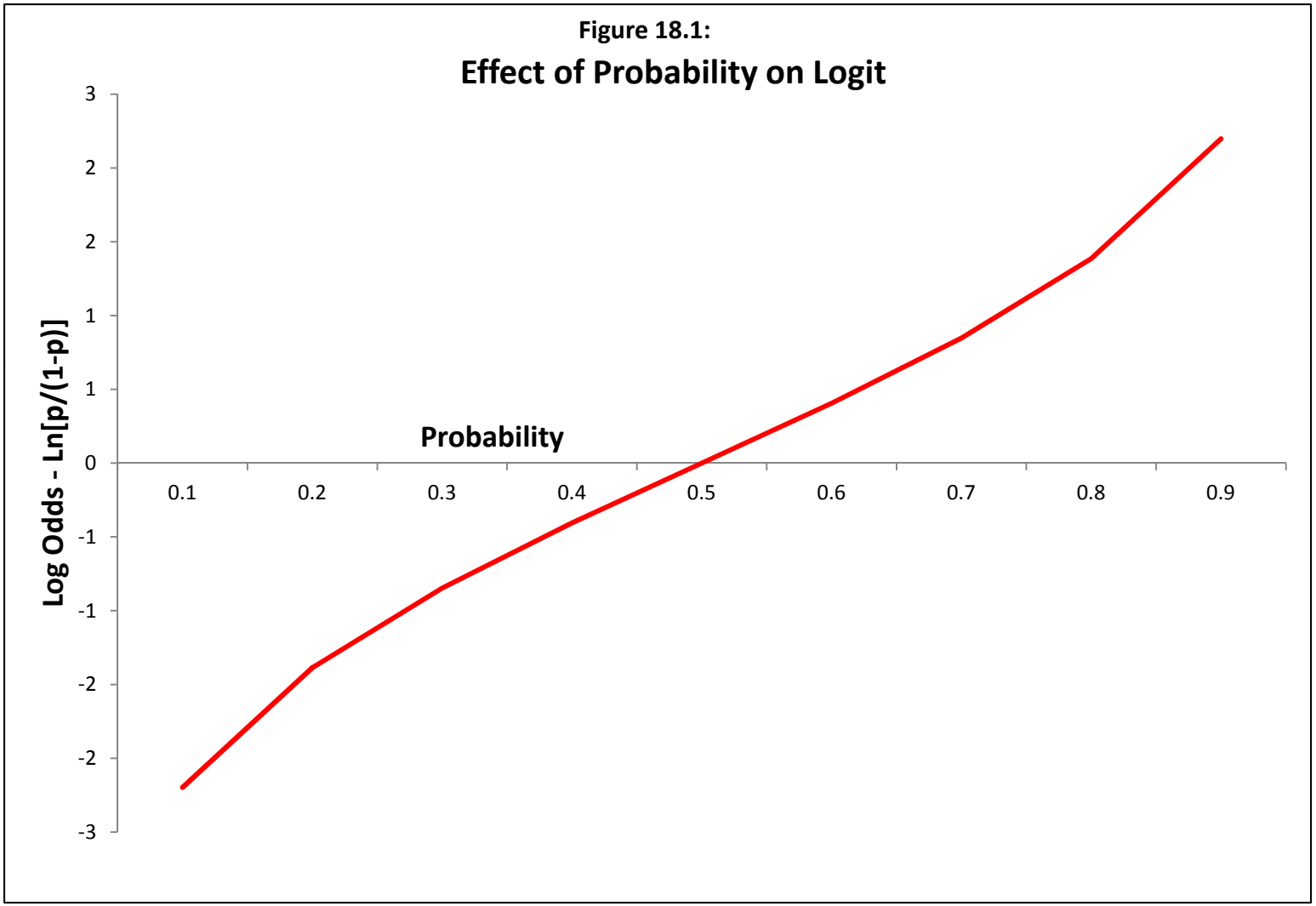
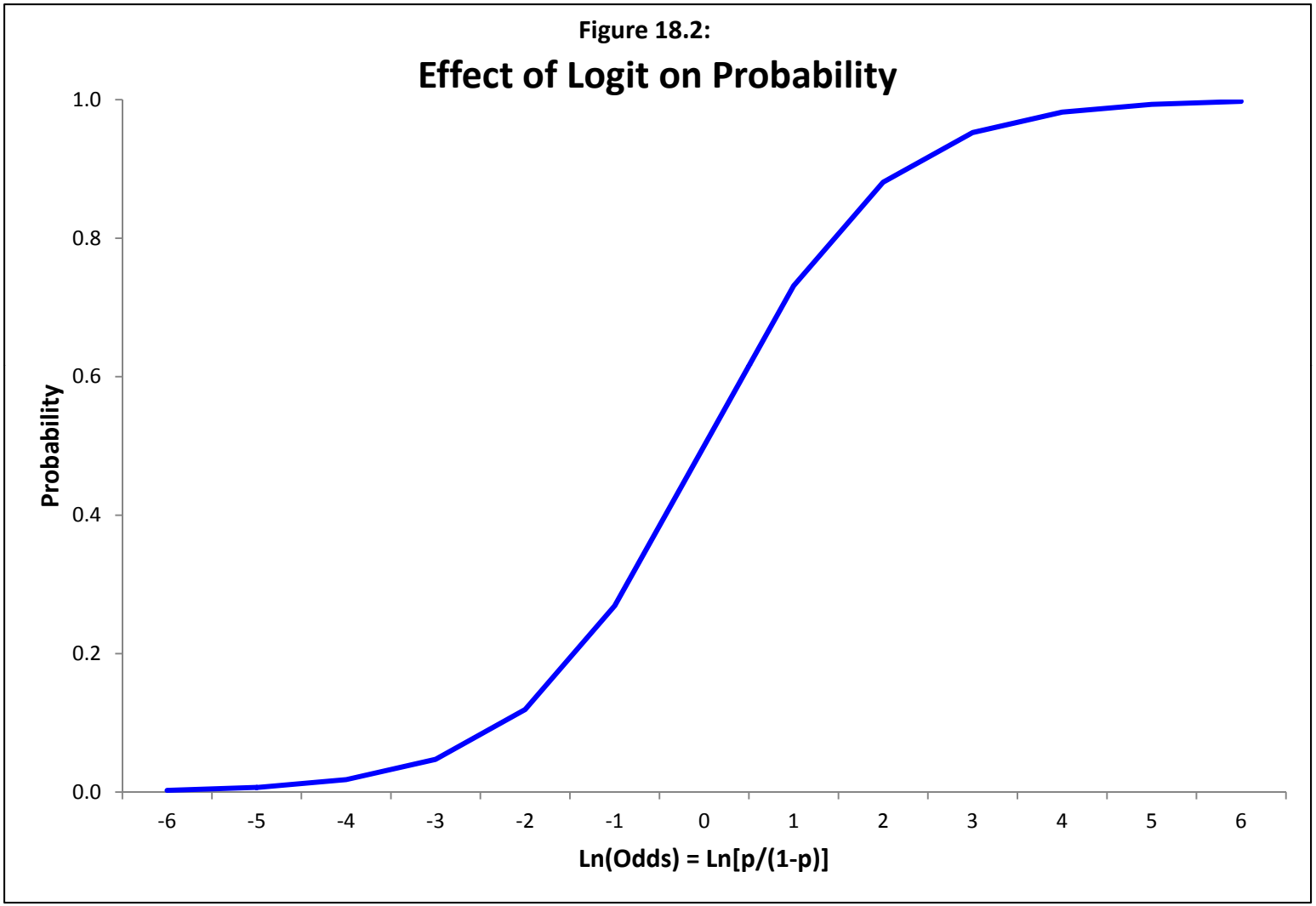

Effect of Probability on Logit
Log Odds - Ln[p/(1-p)]
Probability
3
2
2
1
1
0
-1
-1
-2
-2
-3
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 18.2:


Effect of Logit on Probability
1.0
0.8
0.6
0.4
0.2
0.0
Probability
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Ln(Odds) = Ln[p/(1-p)]

which is a true logistic (S-shaped) function. Some references refer to equation 18.9 as a logit and 18.11 as a logistic. However, they are equivalent functions (Liao, 1994). The probability of a 0 is simply 1 minus the probability of a 1, or

$$P(Y=0) = \frac{1}{1+e^{\beta_0+\sum_1^K \beta_K X_K}} \tag{18.12}$$

As an example, figure 18.3 illustrates the probability that is obtained from a logit model that is estimated by

$$Ln\left[\frac{p}{1-p}\right] = -10 + X$$

where X is a simple variable that varies from 0 to 20. Note the coefficient for X is 1.0. At the low end, the effect of increasing X is minimal in effecting the probability. In the middle, the effect of X is the greatest while at the high end, again, the effect of increasing X on the probability is minimal. This is the nature of a probability function since it is bounded by 0 and 1. The logit simply allows the probability to be regressed against one or more independent variables.

The model is inherently non-linear and must be solved by an iterative method. For the normal logit function, maximum likelihood estimation (MLE) is used. For more complex logit functions, Markov Chain Monte Carlo (MCMC) methods can be used.

Interpretation of the Logit Model

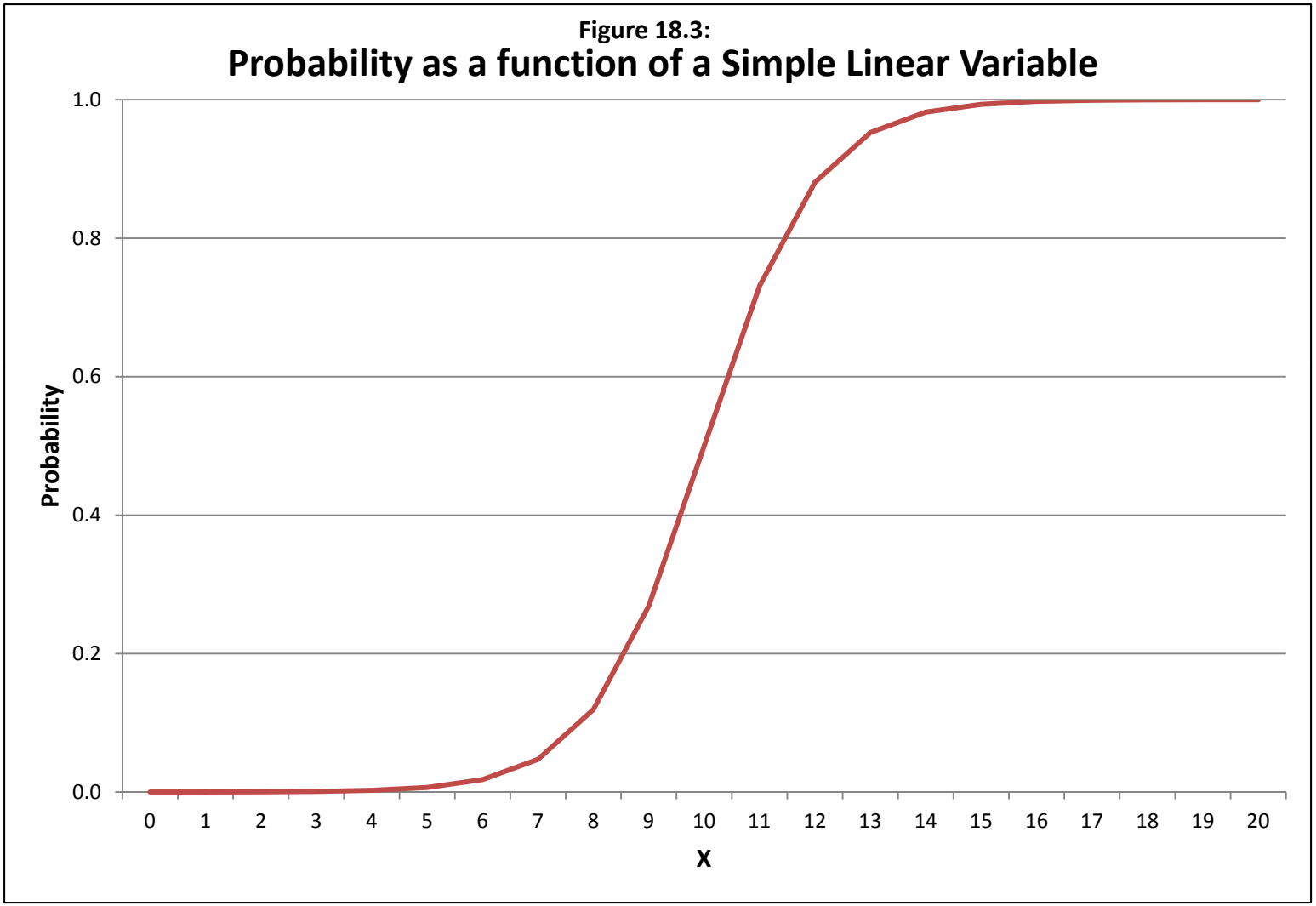
Sign of the Coefficient

Examples will be provided shortly but, there are several ways to interpret the logit model in equation 18.9 (Pampel, 2000). First, there is the sign of the coefficient. As in most regression models, a positive sign indicates that the independent variable increases the probability of the choice being made while a negative sign indicates that the independent variable decreases the probability of the choice being made. Whether we interpret the results in terms of the log of the odds ratio, the odds ratio itself, or the probability, the sign indicates the directional effect of the variable.

Log of the Odds Ratio

Second, there is the log of the odds ratio. Since the model is estimated as a log of the odds function, the interpretation of the coefficients is similar to other regression models, namely the coefficient of each independent variable expresses the change in the dependent variable from

Figure 18.3:


Probability as a function of a Simple Linear Variable
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
X

a one unit change in that variable. However, since the dependent variable is a log of the odds, the coefficients do not have any intuitive meaning in this form other than indicating the sign of the relationship (increasing or decreasing) and the relative strength of the variable as indicated by a Z-test (coefficient divided by standard error).

The use of logged odds for interpretation does have the advantage of symmetry. For example, if the odds of obtaining one alternative (e.g., the odds that a robber will carry some type of weapon) is 9:1 (i.e., the probability of the alternative is 0.9 while the probability of other alternative is 0.1), then the log of the odds for the alternative is 2.1972. For the other alternative, the log of the odds is -2.1972. In other words, the log of the odds of the selected alternative (which takes the value 1) is the opposite of the log of the odds of the non-selected alternative (which takes the value 0).

Odds ratio

Third, a more intuitive way to interpret the logit model is through the odds ratio itself. Equation 18.10 above shows the odds as a function of the exponentiated linear equation. Since the exponent of a sum is equal to the product of the exponent of the parts, we have

$$\left(\frac{p}{1-p}\right) = e^{\beta_0 + \sum_1^K \beta_K X_K} = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} \dots \dots e^{\beta_K X_K + \epsilon} \quad (18.13)$$

The odds ratio can be expressed as the product of the exponentiated coefficients times their variable values and including the error term, ε. In this case, the effect of a unit change in each independent variable on the odds ratio is the exponentiated coefficient. For example, if a coefficient was -0.2, then the effect of a one unit change in that variable on the odds ratio will be $e^{-0.2} = 0.8187$ (or a decreasing effect). Similarly, if a coefficient was 1.1, then the effect of a one unit change in that variable on the odds ratio will be $e^{1.1} = 3.0042$ (or an increasing effect). As mentioned above, the odds ratio has an intuitive meaning in that it indicates the relative likelihood of alternative A versus alternative B.

The percentage change for a one unit increment in the independent variable can be determined by (Pampel, 2000):

$$Percent\ change_K = \left(e^{\beta_K} - 1\right) * 100 \quad (18.14)$$

where β_K is the coefficient of an independent variable in the logit function in equation 18.9 while e^{β_K} is the odds ratio of the variable. To use the example above, if the coefficients was -0.2, then the percentage change from a one unit increase in that variable is -18.1% ($[e^{-.2}-1]*100 = [0.819-1)*100$).

Probability

Fourth, one can express the logit model through a probability itself, essentially solving equation 18.11. The result is a probability function. Unfortunately, the effect of a coefficient on the probability is non-linear and not constant and changes according to the level of the probability. That is, when the probability, *p*, is very low (e.g., 0.1), the effect of an independent variable is also very weak. Similarly, when *p* is very high (e.g., 0.9), the effect of an independent variable is similarly weak. The effects of an independent variable on the probability are strongest when the probability is in the middle range and the absolute strongest when the probability is exactly 0.5.

Variance

Fifth, an important component of a logit model is the variance. With logit models, as with Poisson models, the variance is a function of the mean. That is, the probability, *p*, is the expected value of the distribution:

$$E(Y) = p \tag{18.15}$$

where Y is a binary variable. The variance of a probability is, itself, a function of the mean:

$$Var(Y) = p(1-p) \tag{18.16}$$

This is the similar to the Poisson-based models where the variance of the Poisson is a function of mean and is always underdispersion (variance less than the mean).[1] With ungrouped data, it is not possible for the actual variance to exceed the predicted variance since they are measured exactly the same (McCullagh and Nelder, 1989). With grouped data, however, it is possible for the actual variance to exceed the expected variance. However, since the logit routines in CrimeStat only apply to individual records (i.e., there is no grouping), the variance is always that indicated by equation 18.16.

The Error Term

Finally, let us discuss briefly the error term in the model, ε. In the GLM interpretation (equation 18.1), the error, ε, is the difference between the observed and predicted values. With the OLS model discussed in Chapter 15, the errors are assumed to be normally distributed and

[1] Note that in an Ordinary Least Squares (OLS), the variance is estimated independently of the mean. Thus, there is no confounding of effects. This is one advantage of OLS compared to Poisson or binomial models. On the other hand, OLS does not model skewed distributions very well nor can it model a binary variable.

constant (a condition known as homoscasticity). With the Poisson family of models discussed in Chapters 16 and 17, the errors are normal but not constant (heteroscadstic). For the 'true' Poisson model, they are also Poisson but for the negative binomial model, they are Gamma distributed. We also discussed lognormal error terms in Chapter 17. In all cases, though, the errors are normally distributed.

However, for a probability, the error cannot be normally distributed except in the middle range of the probabilities. Take the example shown above in figure 18.3. At the two extremes – the low end and the high end, the error will be much smaller than in the middle range of the probabilities. In fact, the error will be greatest in the middle. But, also, the errors must be asymmetrical at the two extremes. The closer an estimated probability is to either 0 or to 1, the more likely the errors will be skewed and asymmetrical (meaning that they will fall on one side of the estimate rather than the other. This is just a function of the limits of a probability which have to fall between 0 and 1. In the middle range, however, the errors are generally symmetrical and normally distributed.

McFadden (1973) and Train (2009) have shown that the errors for a logit model are distributed *extreme value* distribution (sometimes called Gumbel or type I extreme value (see also Wikipedia, 2011c). It is part of a family that describes extreme distributions called the Generalized Extreme Value distribution (Wikipedia, 2011d). The extreme value distribution models the maximum or minimum at the extremes of a limited dependent variable, such as a probability. Train (2009) points out that the extreme value gives slightly higher proportions at the extremes of a probability than a normal distribution, and also allow for the asymmetry at the extremes. However, in the middle range, the extreme value distribution is virtually indistinguishable from a normal distribution. It is somewhat similar to a Student's t-distribution though the mathematics is different (Wikipedia, 2011e).

Logit Regression

In CrimeStat, there are three different logit models. One of these is estimated through maximum likelihood (MLE) while the other two are estimated through the Markov Chain Monte Carlo (MCMC) simulation methodology. If readers are unfamiliar with the MCMC method, we suggest that they review Chapters 16 and 17 before going forward in this chapter.

Logit Analysis of Weapon Use for 2007-09 Houston Robberies

MLE Logit

In an MLE logit, the logit model shown in equation 18.9 is estimated with a maximum likelihood estimator. As an example, we use data on 3,709 robberies that occurred within the

City of Houston from 2007-2009. Robberies were selected in which both the crime location and the offender's residence location were known. These came from suspect lists and are only 11% of the total robberies committed within the City for those years. They were selected because the suspect list included information on the age, gender and ethnicity of the offender, whether other suspects were involved, as well as the distance from the residence to the crime location. Additional information on the location of the crime was collected.

The dependent variable was whether a physical weapon had been used, either a firearm, a knife, a stick or another physical object, compared to physical force or the threat of force. Figure 18.4 show the distribution of the robberies and the type of weapon or threat used. Of these 3,709 robberies, 2,333 (or 63%) involved a physical weapon. These were coded as '1' (used a physical weapon) or '0' (did not use a physical weapon). The goal was to estimate the characteristics associated with the use of a physical weapon.

Table 18.1 shows the results of a regression model relating the use of a physical weapon to seven independent variables. The model was estimated with the maximum likelihood (MLE) Logit model in CrimeStat. Only variables that were significant at the $p \leq .05$ or smaller and which had very high tolerances were selected for the model (the process of eliminating non-significant and collinear variables is not shown). See Chapters 15 and 17 for a discussion of multicollinearity.

The log likelihood is substantially negative and the AIC and BIC, statistics used to correct the log likelihood for the number of independent variables (see Chapter 16, p. 16.5) are substantially positive, as would be expected. However, given that there are 3,709 records, we would expect the models to be significant.

Therefore, one has to look at other statistics. In terms of the overall probability, the deviance and the Pearson chi-square are both significant, indicating that the model is significantly different from a random model (which would be expected). On the other hand, when these are adjusted for degrees of freedom (adjusted deviance and adjusted Pearson Chi-square), the statistics are not significant. This indicates that fit of the model was. This is supported by the mean absolute deviation and the measured squared predicted error statistics which shows the model fit quite well (a discussion of these statistics are found in Chapter 16). Keep in mind, though, that the dependent variable is binary which means that there are only values of 0 or 1.

All six independent variables are highly significant. The tolerance statistics indicate that they are almost completely independent (note, this is not surprising since we eliminated collinear statistics while building the model). This is an important point that we keep re-iterating.

Figure 18.4:

Houston Robberies: 2007 to 2009

Location of Robberies by Weapon Use

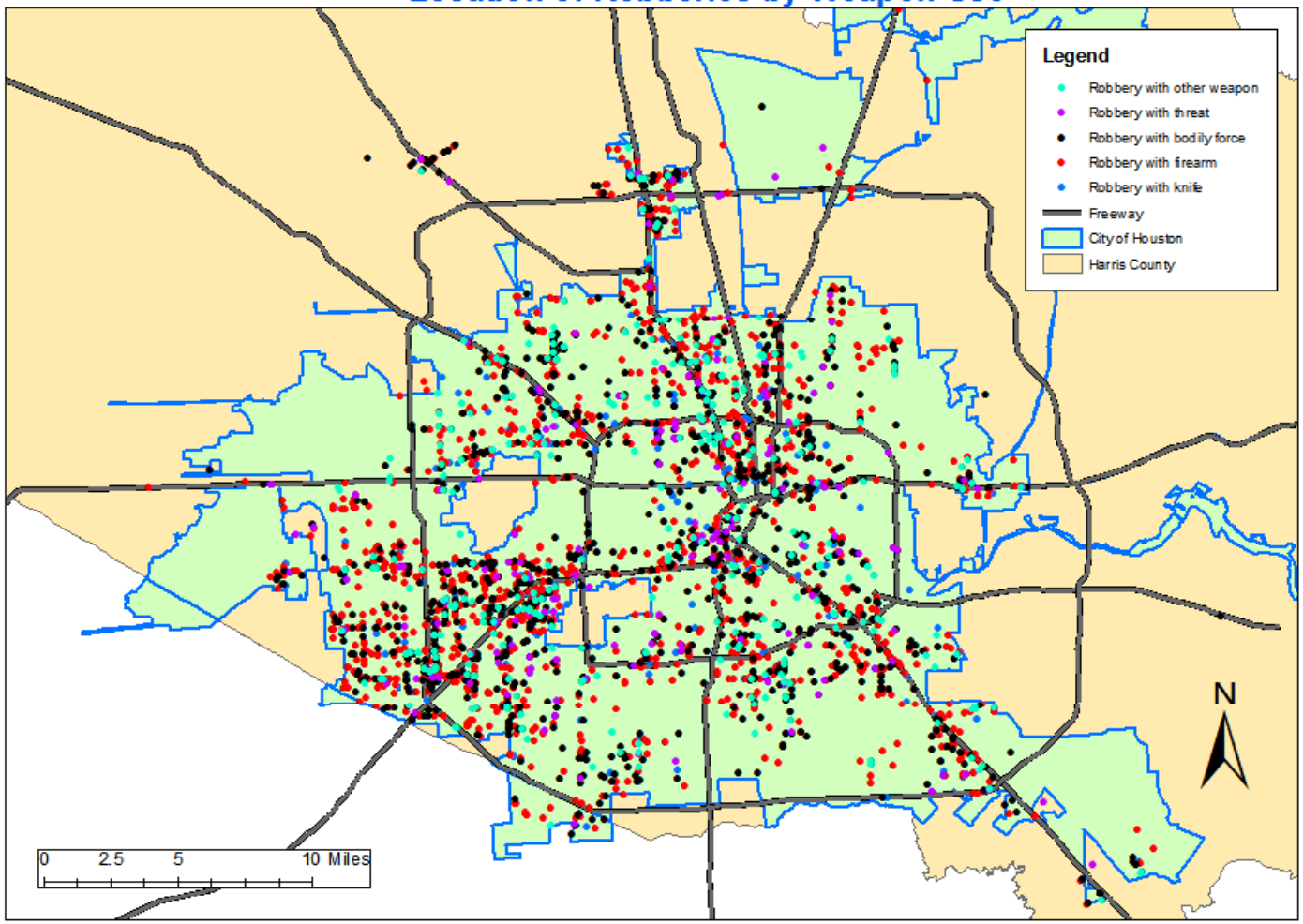


Table 18.1
Weapon Use by 2007-09 Houston Robbers:
MLE Binomial Logit Model

(N=3,709 Robberies with Known Origin & Destination Coordinates)

DepVar:	**WEAPON USE IN ROBBERIES**	
N:	3,709	
Df:	3,696	
Type of regression model:	Logit	
Method of estimation:	Maximum Likelihood	
Likelihood statistics		
Log Likelihood:	-2,345.7	
AIC:	4,707.3	
BIC/SC:	4,757.1	
Deviance:	2,086.1	p: 0.0001
Pearson Chi-Square:	1,373.3	p: 0.0001
Model error estimates		
Mean absolute deviation:	0.4	
1st (highest) quartile:	0.4	
2nd quartile:	0.4	
3rd quartile:	0.5	
4th (lowest) quartile:	0.6	
Mean squared predicted error:	0.2	
1st (highest) quartile:	0.1	
2nd quartile:	0.1	
3rd quartile:	0.3	
4th (lowest) quartile:	0.4	
Dispersion tests		
Adjusted deviance:	0.6	p: n.s.
Adjusted Pearson Chi-Square:	0.4	p: n.s.

Predictor	DF	Coefficient	Stand Error	Tolerance	Z-value	p	Odds ratio
INTERCEPT	1	0.7005	0.147	-	4.76	0.001	2.015
AGE	1	-0.0197	0.003	0.965	-5.67	0.001	0.981
GENDER	1	-0.6059	0.110	0.992	-5.53	0.001	0.546
# SUSPECTS	1	0.2981	0.043	0.979	6.89	0.001	1.347
NIGHT	1	0.5225	0.092	0.985	5.68	0.001	1.686
MEDIAN HOUSEHOLD INCOME	1	-0.000008	0.000002	0.981	-3.47	0.001	1.000
DISTANCE TO DOWNTOWN	1	0.0316	0.007	0.966	4.56	0.001	1.032

Typically, both an MLE and an MCMC model will converge more quickly and will produce cleaner estimates if the independent variables are truly independent.

Examining the effects of the individual variables, younger offenders and those who are male are more likely to use a physical weapon. Looking at the odds ratio of -0.0197 means that for each year of age for a robber, the likelihood of using a physical weapon decreases by about 2% ($[e^{0.0197}-1]*100$). Female robbers (those whose gender value is 1 in the model) are 45% less likely than males to use a physical weapon ($[e^{-0.6059} - 1]*100$).

On the other hand, the more suspects/co-offenders involved in the robbery, the more likely there will be a use of a physical weapon. With an odds ratio of 1.347, each additional co-offender increases the likelihood of using a physical weapon by 35% ($[e^{1.347}-1]*100$) compared to a robbery with only a single offender. Similarly, robberies committed at night time (Midnight to 6 am) are 69% more likely to involve a physical weapon ($[e^{1.686}-1]*100$).

The environmental variables suggest a small effect for income (decreasing) and a small effect for distance (increasing). Why robberies committed farther from downtown involve a greater likelihood of having a physical weapon involved is not clear. For the other variables, the effects are what we would expect.

Note that the odds ratio gives the relative likelihood of the independent variable on the dependent variable. For categorical independent variables, such as GENDER or NIGHT, the comparison is between the group with the value 1 (females and night time respectively) compared to the group with the value 0 (males and other time periods respectively). For continuous independent variables, such as AGE and #SUSPECTS, the odds ratio indicates the incremental effect of a one unit change in that variable.

MCMC Logit

CrimeStat includes both maximum likelihood and MCMC versions of the logit. For comparison, we ran the same model as in Table 18.1 using the MCMC algorithm. There were 25,000 iterations with 5,000 of these being discarded ('burn in'). Hence, the final results were from the 20,000 iterations beyond the 'burn in' sample. Table 18.2 shows the results.

The log likelihood value is stronger (more negative) than for the MLE logit while the AIC and BIC statistics are more positive. The deviance and Pearson chi-square statistics are very similar to the MLE logit and indicate that the model was significantly different than one fit by chance. The MCMC error relative to the standard deviation values are all below 0.05 and the G-R statistics are well below 1.2 (see Chapter 17 for explanation of these indices).

Table 18.2

Weapon Use by 2007-09 Houston Robbers: MCMC Binomial Logit Model

(N=3,709 Robberies with Known Origin & Destination Coordinates)

DepVar:	**WEAPON USE IN ROBBERY**		
N:	3,709		
Df:	3,701		
Type of regression model:	Logit		
Method of estimation:	MCMC		
Number of iterations:	25,000	Burn in: 5,000	
Likelihood statistics			
Log Likelihood:	-2,348.1		
AIC:	4,712.3		
BIC/SC:	4,762.0		
Deviance:	-587.3	p: 0.0001	
Pearson Chi-square:	1,373.6	p: 0.0001	
Model error estimates			
Mean absolute deviation:	0.4		
1st (highest) quartile:	0.3		
2nd quartile:	0.4		
3rd quartile:	0.5		
4th (lowest) quartile:	0.6		
Mean squared predicted error:	0.2		
1st (highest) quartile:	0.1		
2nd quartile:	0.1		
3rd quartile:	0.3		
4th (lowest) quartile:	0.4		
Dispersion tests			
Adjusted deviance:	-0.2	p: n.s.	
Adjusted Pearson Chi-Square:	0.4	p: n.s.	

Predictor	Mean	Std	t-value[p]	.	MC error	MC error/ std	G-R stat	Odds ratio
INTERCEPT	0.6923	0.150	4.60***		0.005	0.035	1.008	1.998
AGE	-0.0197	0.003	-5.65***		0.0001	0.030	1.004	0.981
GENDER	-0.6070	0.110	-5.50***		0.001	0.008	1.000	0.545
# SUSPECTS	0.3005	0.044	6.81***		0.001	0.022	1.003	1.350
NIGHT	0.5249	0.091	5.74***		0.001	0.009	1.000	1.690
MEDIAN HOUSEHOLD INCOME	-0.000008	0.0000	-3.29**		0.0000006	0.024	1.004	1.000
DISTANCE TO DOWNTOWN	0.0318	0.007	4.58***		0.0001	0.011	1.000	1.032

*** p≤.001

** p≤.01

Note, also, that the deviance statistic is negative in Table 18.2. This is because the posterior distribution of the dependent variable (weapon use in robberies) is not normal since it is constrained by the binomial variable to be between 0 and 1 and has a small standard deviation (Spiegelhalter, 2006). Thus, with an MCMC logit model, one might expect a negative deviance. This was not true with the MLE logit model in Table 18.1, however. In either case, the adjusted deviance is not significant, suggesting that the dispersion has been adequately accounted.

The coefficient estimates are almost identical. They differ only in the third decimal place for several values. Similarly, the standard error estimates are also quite similar up through the second decimal place. Finally, the odds ratios are almost identical for the two estimates, up through the second decimal place.

Note that there is no dispersion measure in the logit model. The reason is that the standard deviation of a binomial variable is always:

$$SD_{binomial} = \sqrt{(p)(1-p)} \tag{18.20}$$

In short, the MCMC logit has replicated the MLE logit model for Houston robbery weapon use. So, why run an MCMC model when an MLE will produce almost identical results in a fraction of the time? The reason has to do with running more complex models than a simple logit, particularly a binomial logit with an estimate of spatial autocorrelation. Chapter 19 will discuss that issue.

MCMC Logit-CAR/SAR

The final logit model is a spatial model. This will be discussed in Chapter 19.

Probit Model

MLE Probit

The logit is the most commonly used way to model a binary variable. But, there are other functions that can also linearize a binary dependent variable. One commonly used one is the probit function for which the link function was defined in equation 18.5. The probit expresses the inverse of the cumulative standard normal distribution as a linear function of independent variables (without an error term):

$$p(Y=1) = \Phi^{-1}(p_i) = \beta_0 + \sum_1^K \beta_K X_K \tag{18.21}$$

where Φ is the cumulative standard normal distribution,

$$\Phi(\mathbf{x}_i^T\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_i^T\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \tag{18.22}$$

The inverse of the cumulative standard normal distribution is a Z-score and, essentially, the probit is a cumulative Z-score for a one-tailed probability:

$$p(Y = 1) = \Phi\{\beta_0 + \sum_1^K \beta_K X_K\} \tag{18.23}$$

The area under the standard normal distribution is 1.0. Starting at minus infinity, the area under the curve can be expressed as a probability and the link function, η, is a linear regression of the Z score of the event probability (Liao, 1994). The probability of a non-event is 1 minus the probability, or

$$p(Y = 0) = 1 - \Phi\{\beta_0 + \sum_1^K \beta_K X_K\} \tag{18.24}$$

Interpreting the coefficients is not intuitive because it involves additive effects of the intercept and independent variables on the inverse of the cumulative standard normal distribution. Also, unlike the logit function, there is not an odds ratio. Nevertheless, the signs of the coefficients are in the same direction as for the logit model and the Z-values produced by coefficients divided by their standard errors are usually of the same magnitude.

To see this, we model weapon use among the Houston robbers (Table 18.3). Comparing this table with MLE logit model (Table 18.1), the likelihood statistics are virtually the same; the signs of the coefficients are identical and the Z-scores of the coefficients are of the same magnitude. The values of the coefficients are, of course, very different since they express the dependent binary variable in different units. The model is estimated in CrimeStat with maximum likelihood. At this point, there are no MCMC probit models though we may add them in later versions.

Utility of the Probit Model

With most datasets, the logit and probit models will produce almost identical conclusions. They differ primarily in the tails of the distribution with the probit approaching the limiting ends of the probability more quickly than the logit.

Table 18.3:
Weapon Use by 2007-09 Houston Robbers:
MLE Probit Model

(N=3,709 Robberies with Known Origin & Destination Coordinates)

DepVar:	**WEAPON USE IN ROBBERY**	
N:	3,709	
Df:	3,696	
Type of regression model:	Probit	
Method of estimation:	Maximum Likelihood	
Likelihood statistics		
Log Likelihood:	-2,347.4	
AIC:	4,710.9	
BIC/SC:	4,760.6	
Deviance:	4,479.6	p: 0.0001
Pearson Chi-Square:	2,472.9	p: 0.0001
Model error estimates		
Mean absolute deviation:	0.9	
1st (highest) quartile:	0.6	
2nd quartile:	0.6	
3rd quartile:	0.9	
4th (lowest) quartile:	1.4	
Mean squared predicted error:	1.2	
1st (highest) quartile:	0.6	
2nd quartile:	0.6	
3rd quartile:	1.2	
4th (lowest) quartile:	2.2	
Dispersion tests		
Adjusted deviance:	1.2	p: n.s.
Adjusted Pearson Chi-Square:	0.7	p: n.s.

Predictor	DF	Coefficient	Stand Error	Tolerance	Z-value	p
INTERCEPT	1	0.4550	0.089	-	5.10	0.001
AGE	1	-0.0121	0.002	0.965	-5.68	0.001
GENDER	1	-0.3706	0.068	0.992	-5.47	0.001
# SUSPECTS	1	0.1656	0.024	0.979	6.89	0.001
NIGHT	1	0.3181	0.055	0.985	5.78	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.000005	0.000001	0.981	-3.38	0.001
DISTANCE TO DOWNTOWN	1	0.0191	0.004	0.966	4.63	0.001

Using the example discussed in chapters 15, 16 and 17, we model 2006 Houston burglaries in 1,179 traffic analysis zones (TAZ). But, instead of modeling the number of burglaries per TAZ, we created a binomial variable for one or more burglaries. The dependent variable was whether the TAZ had one or more burglaries in 2006 and the two independent variables were the number of households in 2006 and the 2000 median household income. Table 18.4 shows the result of the probit model while table 18.5 shows the result of the logit model.

There are some subtle differences. The logit model has a higher log likelihood value (i.e., less negative) and lower AIC and BIC values, suggesting that it is a better probability model. The model error statistics (mean absolute deviation and mean squared predicted error) are similar though the logit does a better job in fitting the fourth (lowest) quartile.

The coefficients, however, are a little different. The intercept for the logit is significant while that of the probit is not. The coefficient for median household income is almost significant in the logit model ($p \leq 0.1$) while it not significant in the probit model. Whether these differences are meaningful would depend on what the researcher is willing to assume. As mentioned, the probit assumes an underlying normal distribution while the logit does not. If the transition from a measured null response (0) to a counted response (1) is assumed to be gradual, then the probit may make more theoretical sense.

Figure 18.5 graphs the results of the two models. As seen, the probit model levels off more quickly than the logit model. That is, at the low end, it approaches both the low and high asymptote more quickly than the logit. The probit shows a more gradual change than the logit, which could be a more realistic representation of the shift in probabilities from the null condition to the prevalence of the phenomenon.

Nevertheless, the two models are very highly correlated. Hahn and Soyer (2005) make the point that the two models will be different if the values at the ends are of interest. For most other tests, however, the estimated probabilities will be very similar.

Conclusion

We have examined two different models for estimating the effects of independent variables on a binary dependent variable, the logit and the probit. The logit is clearly more convenient to use given that the exponentiated coefficients can be expressed in terms of the odds ratio. That is the main reason that it more widely used. In Chapter 19, we will show how an MCMC version of the logit can be adapted to estimate spatial autocorrelation in the dependent variable.

Figure 18.5:

Logit and Probit Predictions of Houston Burglaries: 2005-2007

1,179 Traffic Analysis Zones with One or More Burglaries

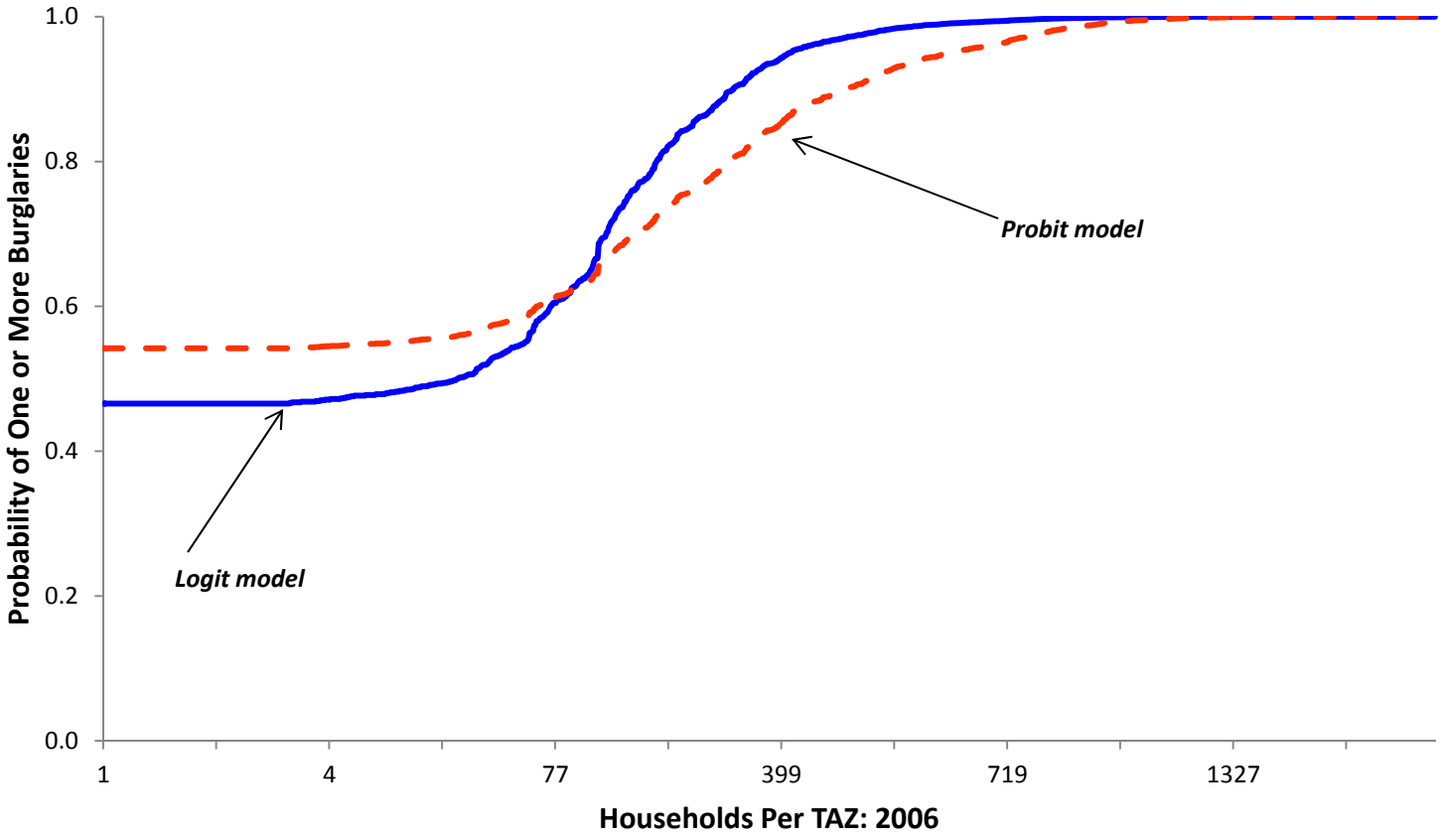


Table 18.4:

Predicting Burglaries in the City of Houston: 2006

MLE Probit Model

(N= 1,179 Traffic Analysis Zones)

DepVar:	**ONE OR MORE BURGLARIES**	
N:	1,179	
Df:	1,175	
Type of regression model:	Probit	
Method of estimation:	Maximum Likelihood	
Likelihood statistics		
Log Likelihood:	-427.5	
AIC:	863.0	
BIC/SC:	883.3	
Deviance:	347.4	p: 0.0001
Pearson Chi-Square:	220.4	p: 0.0001
Model error estimates		
Mean absolute deviation:	0.2	
1st (highest) quartile:	0.1	
2nd quartile:	0.2	
3rd quartile:	0.1	
4th (lowest) quartile:	1.5	
Mean squared predicted error:	0.1	
1st (highest) quartile:	0.0	
2nd quartile:	0.1	
3rd quartile:	0.0	
4th (lowest) quartile:	0.3	
Dispersion tests		
Adjusted deviance:	0.3	p: n.s.
Adjusted Pearson Chi-Square:	0.2	p: n.s.

Predictor	DF	Coefficient	Stand Error	Tolerance	t-value	p
INTERCEPT	1	0.0252	0.083	-	0.03	n.s.
HOUSEHOLDS	1	0.0023	0.0002	0.994	14.34	0.001
MEDIAN HOUSEHOLD INCOME	1	0.000002	0.000002	0.994	1.28	n.s.

Table 18.5:

Predicting Burglaries in the City of Houston: 2006

MLE Logit Model

(N= 1,179 Traffic Analysis Zones)

DepVar:	**ONE OR MORE BURGLARIES**	
N:	1,179	
Df:	1,175	
Type of regression model:	Probit	
Method of estimation:	Maximum Likelihood	
Likelihood statistics		
Log Likelihood:	-389.8	
AIC:	787.7	
BIC/SC:	807.9	
Deviance:	325.4	p: 0.0001
Pearson Chi-Square:	222.6	p: 0.0001
Model error estimates		
Mean absolute deviation:	0.2	
1st (highest) quartile:	0.1	
2nd quartile:	0.2	
3rd quartile:	0.1	
4th (lowest) quartile:	0.4	
Mean squared predicted error:	0.1	
1st (highest) quartile:	0.0	
2nd quartile:	0.1	
3rd quartile:	0.0	
4th (lowest) quartile:	0.2	
Dispersion tests		
Adjusted deviance:	0.3	p: n.s.
Adjusted Pearson Chi-Square:	0.2	p: n.s.

Predictor	DF	Coefficient	Stand Error	Tolerance	t-value	p
INTERCEPT	1	-0.3591	0.151	-	-2.38	0.05
HOUSEHOLDS	1	0.0073	0.001	0.994	10.31	0.001
MEDIAN HOUSEHOLD INCOME	1	0.000006	0.000003	0.994	1.88	n.s.

On the other hand, the probit model has applicability in *random utility* theory which will be discussed in Chapter 21. Train (2009) argues that the probit model can allow for variations in the 'tastes' of decision makers whereas the logit model imposes greater restrictions on the interpretation of coefficients. It can be used to estimate non-constant error variance (heteroscedastic probit models; see Train, 2009) while the logit cannot. But, in general, there really is not much of a difference in their conclusions when applied to the same data.

The final point is that a binary variable, whether measured by the logit or the probit model, is the simplest form of modeling a choice made by a decision-maker. Hence, the logit form (and to a lesser extent, the probit) has widespread applicability in decision theory and is the basis of discrete choice modeling (Train, 2009; McFadden, 1973). Chapter 21 will discuss this.

References

Chen, G. & Tsurumi, H. (2011). Probit and logit model selection. *Communications in Statistics – Theory and Methods*, 40, 159-175.

Greene, W. H. (2008). *Econometric Analysis* (sixth edition). Pearson Prentice Hall: Upper Saddle River, NJ.

Hahn, E. D. & Soyer, R. (2005). Probit and logit models: Differences in the multivariate realm. Unpublished paper. http://home.gwu.edu/~soyer/mv1h.pdf.

Hosmer, D. W. & Lemeshow, S. (2001). *Applied Logistic Regression: Textbook and Solutions Manual*. Wiley-Interscience, J. Wiley & Sons: New York.

Lambert, D. & Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *J. Amer. Stat. Assoc.*, 90, 1225-36.

Liao, T. F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. Sage University Paper 101, Sage Publications, Inc: Thousand Oaks, CA.

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention*, Vol. 37 (1), 35-46

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed). Chapman & Hall: London.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior, in Zarembka, P. (ed.), *Frontiers in Econometrics*, Academic: New York.

Pampel, F. C. (2000). *Logistic Regression: A Primer*. Sage University Paper 132, Sage Publications, Inc.: Thousand Oaks, CA.

Spiegelhalter, D. (2006). Some DIC slides. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml

Train, K. (2009). *Discrete Choice Methods with Simulation* (2nd edition). Cambridge University Press: Cambridge.

References (continued)

Wikipedia (2011a). Binomial probability. *Wikipedia*, http://en.wikipedia.org/wiki/Binomial_probability.

Wikipedia (2011b). Jacob Bernoulli. *Wikipedia*. http://en.wikipedia.org/wiki/Jacob_Bernoulli.

Wikipedia (2011c). Gumbel distribution. *Wikipedia*, http://en.wikipedia.org/wiki/Gumbel_distribution.

Wikipedia (2011d). Generalized extreme value distribution. *Wikipedia*, http://en.wikipedia.org/wiki/Generalized_extreme_value_distribution.

Wikipedia (2011e). Student's t-distribution. *Wikipedia*. http://en.wikipedia.org/wiki/Student%27s_t-distribution.

Wikipedia (2011f). Overdispersion. *Wikipedia*. http://en.wikipedia.org/wiki/Overdispersion